I see it in your eyes : Learnings of the world's shallowest CNN recognising emotions from muted in-the-wild video-chats real-time

Vedhas Pandit[1], Maximilian Schmitt[1], Nicholas Cummins[1], Björn Schuller[1,2]

Abstract

A robust value- and time-continuous emotion recognition is of huge relevance to healthcare. A real-time patient monitoring sytem understanding the patient's physical and emotional state can help doctors make an appropriate diagnosis and treatment plan. In-person and web-assisted patient-patron conversations are beneficial likewise. Presence and absence of assistive technology, that can detect cues a human could miss, can be the difference between life and death.

For the first time ever, we present the shallowest realisable CNN consisting of a single filter in its only intermediate 1D convolutional layer; the learnings of which we prove are generalisable to german, hungarian and chinese populations. We draw insights from visualisations of the trained filter weights and the facial action unit (FAU) activations of the subjects featured in the in-the-wild, spontaneous video-chat sessions of the SEWA corpus, i. e., the inputs to the model. The presented cross-cultural performance is a testimony to the universality of FAUs in expression and understanding of the human affective behaviours. These learnings were found to be consistent with the human perception of emotion expression. Because the FAUs can be extracted real-time using openFACE-like tools, and because the models we designed are shallow, this study proposes a promising architecture for a robust, end-to-end, value- and time-continuous, pan-cultural, real-time affect prediction.

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]GLAM – Group on Language, Audio & Music, Imperial College London, UK

Preprint submitted to Journal of Information Processing and ManagementNovember 11, 2019

Keywords: Feature relevance, in the wild, real-time, emotion recognition, model explainability, layerwise relevance propagation, pattern attribution.

1. Introduction

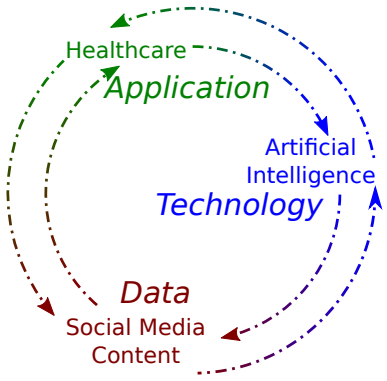

Figure 1: Synergy between social media content (the data), artificial intelligence (the technology), and healthcare – an application end of the utmost importance.

Importance of social media environments as a new source of data for healthcare cannot be overstated today. The potential for the health care industry to leverage existing social media data for better quality, cost, and quality measurement is unparalleled. Likewise, deep learning advancements have made number-crunching and information extraction possible at a scale that is unprecedented. This simultaneous growth has opened doors to many next-generation applications that can utilise social media data and deep learning for a better healthcare.

1.1. Social Media Analytics and Healthcare

Innovative approaches to building social data-driven health information can help create better healthcare decisions across the entire healthcare domain – for patients, providers, and health systems [1]. Because the social media content gives us a sneak peek into the lives, emotions and opinions of the real people like never before, we can leverage this data to recognise patterns that relate to their health needs. One can gain insights into habits and preferences of the

individuals, review which treatments work the best for them and which ones don't – as a consequence of their habits and preferences, or other factors such as ethnicity, age, and gender. Such information can help health professionals gain a much better understanding of the treatments and patient-education that a population is most receptive to, and adapt accordingly [2]. Social media data are also valuable for businesses. It can help those creating the marketing and social elements of a social media strategy; improving engagement with the potential customers [3, 4]. Likewise, social data analytics is a powerful way to discover people that will likely lead a fitness regime, and to also recognise the precise stimulus to keep them motivated [5, 6]. Social media platforms enable patients to share their personal experience with a particular product or a service to a large audience [7]. This can help healthcare professionals to identify the gaps in the system, which in turn can, for example, help reduce the waiting times, improve the customer satisfaction, doctor-patient relationship, likely delivering also a better health outcome [8]. Web-assisted conversations, e. g., Facebook Live, or personalised Skype Calls, can be put to use to convey and address grievances, patient counselling and consultations [9, 10].

1.2. AI and Social Data: The unique challenges and the symbiosis

1.2.1. AI Today

Simultaneous to rise in social media content, the progress in the data mining technology has been stupendous. From convincingly mimicking the human speech and the individual voices with a minimal training data [11–14], to generating high resolution 'deep-fake' videos [15, 16], to penning down a highly coherent synthetic text on any given topic [17–19]; the advancements in AI are staggering. Such is the velocity of this growth that the finesse with which AI performs these tasks remains hard to believe, while ironically, these capabilities and their usage across the board is becoming a common knowledge day by day. For example, to put to the test the prowess of the technology, the introductory paragraphs of this manuscript were partly generated using the very recently released GPT-2 language model [17]. GPT-2 model is claimed to have learnt

the English language (specifically, the language model) on its own by crawling through dataset[3] featuring text from 45 million links [17]. What is frighteningly impressive about this human-like coherency is that the GPT-2 model is able to provide large number of auto-completion suggestions, with varying approaches and opinions relating to the same topic, all the while remaining consistent with the writing style and mannerism of the provided input.

Advancements in machine learning – particularly in the deep learning domain theory and frameworks – has been a primary driving factor to this astounding progress. Much of the AI advancements are also due to general awareness and recent influx of the annotated datasets in the public domain. While the traditional machine learning methods necessitate explicit feature extraction step, the deep learning models can consume even the raw inputs directly. Furthermore, the trained models can be stacked together to realise an end-to-end system accomplishing more complex tasks, in a relatively short amount of time [20]. In the case of the deep learning models, the learnings of the trained models need not be lost. The learnings serve as a better starting point for the training of the model further on a related or a completely different task, in a process called transfer-learning. The inherent modularity and retainability of the deep learning architecture has helped the modern machine learning models cope with the scalability challenges posed by the big data, better than their traditional counterparts such as the linear discriminant analysis, random forest, support vector machines [21].

1.2.2. AI-driven Data Augmentation, aka DeepFakes

While advancements of such a scale are a boon in certain sense – enhancing predictive power and utility of this technology; these advancements also present some unique challenges. Because this new paradim of learning is 'example-driven' (i. e., learning from the examples directly, as opposed to conventional rule-based learning), it is reasonable to anticipate rise in better generalising,

[3] WebText dataset: `https://github.com/openai/gpt-2-output-dataset`

more meaningful data augmentation strategies. While this is good for advancing AI technology as a whole, we can expect to see surge in synthesised social media content likewise. As an unfortunate consequence, a corruption of social media content coming in from real humans for malicious purposes is inevitable. The technology in the wrong hands can be very devastating. It could easily trick the companies and individuals to send scammers the money, hack the political discourse, weaken the democracies [22]. While stronger and smarter AI discriminators are being proposed and trained, the AI-based generative models too are getting more deceptive and more powerful than ever, day by day. The boundaries between the synthesised and the real data have slowly begun to disappear.

1.2.3. Data Privacy and AI

Data anonymisation and privacy is especially crucial in the fields such as healthcare, where the sensitive data (e. g., patient's vital signs) is the essential component and driving force of the system. AI-driven data augmentation makes it easier to generate synthetic data samples encapsulating essential and realistic information, with simultaneous anonymisation for data privacy. On the other side, unauthorised, unconsented use of the collected data remains a concern for continuous digitalization of healthcare systems. As a first step, European General Data Protection Regulation (GDPR) has brough forward regulatory directives, for user consent provisioning and for coordination across data services. A number of AI and block-chain driven approaches too have been proposed for a more secure health-information exchange [23–25]. Ironically, AI also presents itself as a challenge to the data privacy issue, thanks to advancements researchers have made in the field of re-identification of patients from the anonymised data [26–28]

1.2.4. Decision Explainability

The primary and the earliest criticism of deep learning methodology (continuing up to even until very recent times) has long been its inexplicability.

However, there have been monumental developments in the field of 'model explainability' recently [29–39]. Model explainability is especially crucial for a field like healthcare, where every diagnosis and treatment needs to be based on complete understanding of the available information; where an incorrect treatment can prove to be fatal.

1.3. Emotion-aware AI and Healthcare

Data mining, i. e., creating artificial intelligence based application has helped improve diagnosis and treatment of various diseases, e. g., diagnosis of cancer [40], sleep apnea [41, 42], diabetes [43, 44], cardiovascular diseases [45], and assessment of psychological health [46, 47]. The technology, therefore, could be used to identify 'red flag' situations in real-time, which could then be avoided. For example, we could identify people prone to depression or suicide, or those who had negative and potentially life-threatening experiences in the past to help guide them through the challenges of living with a chronic illness - helping to make it an easier transition. Likewise, the technology has been proven to be able to identify other mental conditions, e. g., bipolar disorder [48], autism [49], depression [50]. As discussed in the introductory paragraph, an assistive technology recording dyadic conversations and estimating psychological state of the recorded people can be envisioned.

However, such a system needs to meet very high standards of requirements if it was to be used in a crucial sector like healthcare. First, it needs to be robust enough to be used on non-laboratory, unconstrained, noisy, i. e., 'in-the-wild' data – featuring spontaneous behaviours. It should ideally be able to capture nuances of emotions/affect, without resorting to a crude classification into three (positive, negative and neutral) or six basic emotion classes. That is, value-continuous multidimensional affect regression is preferred. It should be ideally able to track one's emotions continuously in time. The prediction model should ideally be also fast enough to able to run in real-time. Fortunately, recent advancements in machine intelligence, particularly for the affective computing field, are steady steps towards making all these goals within our reach in the

near future.

1.3.1. Scope of This Paper

As discussed earlier, decision explainability is especially crucial for an application domains such as healthcare, where people's lives are at stake. In this paper, we unearth learnings of a shallowest possible CNN one can ever realise, as it learns to predict human emotions from in-the-wild videos. The training and predictions are of value- and time-continuous affect dimensions of people coming from very different cultures, contexts, gender and age-groups, without making use of any audio or textual data.

In Section 2, we briefly discuss research that is directly relevant to our current study. In the sections next, we discuss in detail our dataset (Section 3), including the insights we gained from the statistical analysis we did on the features and the labels contained in the dataset. We then explain the overall experimental design pipeline, different types of models we experimented with, how the powerful yet the shallowest-realisable CNN evolved through these experiments (Section 4). In Section 5, we also summarise insights we gained by analysing the trained weights of the model, how the input features map to the output labels. In Section 6, we summarise our findings, briefly mentioning also limitations of our study. We put forward various avenues for the future work as the logical next step – concluding with some of the research paths we have already begun venturing into.

2. Related Research

We discuss next the related research so far by various research groups across the globe, including our own. The target research problem here is the robust value- and time-continuous recognition of affective dimensions (e. g., arousal and valence) on in-the-wild audiovisual recordings, featuring spontaneous affective behaviours in the conversational context. While there exist many databases already featuring labelled affective behaviours (e. g., GENEVA[51], SMARTKOM[52], VAM-faces[53], SEMAINE[54], AVEC'13[55], Belfast induced 1[56], Belfast induced 2[56], CCDb[57], RECOLA[58], AVEC'14[59], MAHNOB Mimicry[60],

4D CCDb[61], SEWA[62], GRAS^2[63, 64]), each of these databases either features prompted non-spontaneous behaviours, and/or has been recorded in the laboratory settings (thus, is not in-the-wild), and/or the featured labels are categorical (thus, missing out on the nuances of the affect), and/or the labels featured are at the sample level (thus, missing out on the temporal trends in the affect), and/or are too small [62]. The only exceptions are SEWA and GRAS^2 which have been recorded in the wild, feature recorded audiovisual data with real-life noise as part of every recorded modality, labelled with value- and time-continuous annotations for arousal and valence dimensions by multiple annotators. The GRAS^2 database has a further advantage that the recordings are arguably more Gabor effect-free. A model trained on the more challenging, more noisy, and more in-the-wild corpus such as GRAS^2 has been proven to generalise better than a model trained on a data recorded with more constraints [65]. However, while the GRAS^2 database is challenging and therefore ideal, the time-bound consent to use it for research purposes has now expired.

On the other hand, the 'Automatic Sentiment Analysis in the Wild' (SEWA) corpus [62] has consistently featured in the 'Affect Recognition' sub-challenge of the Audio/Visual Emotion Challenge and Workshops since a few years (AVEC 2017, 2018, 2019) [66–68]. It is arguably the most popular in the wild public database available to date, thoroughly investigated by research groups from across the globe, that features time-continuous, high resolution labels for multiple dimensions of affect. The participants of the AVEC challenges [69–77] compete against one another to correctly predict arousal, valence and liking dimensions featured in the dataset across different cultures, based on the audio, textual and video features provided in the dataset. The bag-of-words representation of these features computed using openXBOW [78] has been shown to perform well across all the three modalities.

In this paper, we combine and extend two of our previous research works. For the very first time, we ventured into the territory of 'explainable AI' for time- and value-continuous in the wild affect predictions [79]. To investigate how the trained models use the text modality to deliver decent performance across

all of the three featured affect dimensions, we computed feature attributions of the individual 521 features. These 521 features effectively represent frequency of specific 521 words in a 6 second window. Action markers such as <laughter> and <slight_laughter> are highly informative when predicting arousal and valence, and the model too was found to utilise the corresponding derivative counts the most. Even more interestingly, the model reassigned high relevance to words implying contexts when predicting liking, e. g., 'dazu'/'außerdem' (therefore), 'endlich' (at last), 'über' (over (something)), 'zusammenhang' (context), 'weil' (because) consistent to human perception. We also established that the text-features themselves were informative enough that a simple feedforward network was able to utilise these features effectively and predict the arousal, valence and liking dimensions, for the German participants.

In another research [80], we questioned whether it is imperative to have recurrent neural network (RNN) for a value-continuous time-series prediction problem – specifically, for an affect prediction problem. While in theory, the RNNs can handle long-term dependencies, there are fundamental reasons why they fail to in practise [81, 82]. RNNs featuring nodes with internal memory units (e. g., Gated Recurrent Units (GRU) [83] and Long Short Term Memory (LSTM) [82]) attempt to resolve this exact problem. However, despite the invention of memory units, RNNs are not exactly problem-free [84, 85]. RNNs continue to suffer from vanishing gradient and exploding gradient problem [86]. Convolutional neural network (CNN), on the other hand, has a further advantage that it allows for data parallelisation. CNNS, thus, can be effectively way faster to train and to test [21]. When creatively used, CNNs can be used to model input to output mapping for the time-series data [87]. We trained and compared performance of an LSTM-RNN model and two CNN models with similar topologies. The models were trained and tested using the audio features of the AVEC'18 Continuous Emotion Recognition Subchallenge data, subset of AVEC'19 challenge. The models were, thus, effective cross-culturally. They were trained on the data obtained from the German-speaking subject, and were tested on Hungarian-speaking subjects We established that the CNNs were as

Table 1: Participant count and duration for the data splits of the SEWA dataset for AVEC'19 challenge [68], and for the experiments featured in this paper

Culture	**Partition**	**#Subjects**	**Duration (mmm : ss)**
German	Training	34	093 : 12
German	Development	14	037 : 46
German	Test	16	046 : 38
Hungarian	Test	66	133 : 12
Chinese	Test	70	200 : 46

effective as the RNNs, if not more, when predicting the arousal and valence dimensions.

In the current research work, we endeavour to make use of the video modality to predict the arousal and valence dimensions using CNN models similar to those we used in [80].

3 and 4 are must haves!

3. The Dataset

As discussed earlier, for the lack of as cross-cultural, and as in-the-wild dyadic conversational data, we use the SEWA dataset for this study [62]. SEWA happens to be also a very popular dataset, available in the public domain for free, on which many benchmarking studies have been performed so far by various research groups from across the globe. This dataset has previously been used in various audiovisual emotion recognition challenges (e. g., AVEC 2017, AVEC 2018, AVEC 2019) [66–68]. The dataset features spontaneous dyadic conversations over the internet (i. e., video chat sessions) between participants, where they discuss an advertisement shown to them right before their discussions begin. The participants came from different age-groups, with various degrees of mutual acquaintance, different cultural backgrounds. They conversed in one of

these six languages; Chinese, English, German, Greek, Hungarian, and Serbian. With 201 male and 197 female participants (male / female gender ratio = 1.020) and with about 60 pairs from each of the 6 linguistic populations, the the dataset features behavioural data from diversified and balanced demographics. So far, the audiovisual recordings of only Chinese, German and Hungarian participants have been fully annotated and are available for further research (cf. Section 3).

3.1. Overview of the labels and the features contained in the dataset

SEWA dataset features value- and time-continuous annotations for arousal, valence and liking (how much participants liked the advertisement), annotated by five to six annotators coming from the same linguistic background as of the participants of the audiovisuals they annotate. While annotating, therefore, the annotators can make use of every data modality available to them, namely the linguistic content (i. e., what is being spoken), the audio content (i. e., how it is being spoken, e. g., prosody, pitch, tone of the voice), and the visual content (i. e., non-verbal cues, facial expressions).

Likewise, to train a machine learning algorithm, in addition to audiovisual recordings and transcriptions of the conversations, the dataset contains various other linguistic, audio and video features, e. g., Mel Frequency Cepstral Coefficients (MFCCs), their 1^{st} and 2^{nd} derivatives, eGeMAPS audio feature set [88] extracted using openSMILE [89, 90], raw locations of the facial landmark points, facial action unit (FAU) features extracted using openFACE[4] [91].

[4] https://github.com/TadasBaltrusaitis/OpenFace

Table 2: Illustrations of what the FAU activations look like.
(Image credits: With permission from `imotions.com/blog/facial-action-coding-system/`)

Action Unit (AU)	Description	Example Movement From	To
1	Inner Brow Raiser		
2	Outer Brow Raiser		
4	Brow Lowerer		
5	Upper Lid Raiser		
6	Cheek Raiser		
7	Lid Tightener		
9	Nose Wrinkler		
10	Upper Lip Raiser		
12	Lip Corner Puller		
14	Dimpler		
15	Lip Corner Depressor		
17	Chin Raiser		
20	Lip stretcher		
23	Lip Tightener		
25	Lips part		
26	Jaw Drop		
45	Blink		

3.2. FAU Activation-Label Statistics and Insights

In this study, we restrict our attention to employing FAU features, in conjunction with speaker-turn activation feature – which is indicative of at least one of the two participants speaking. An FAU is defined by Facial Action Coding System (FACS) – a predefined set of different simultaneous facial muscle movements. In the experiments next Table 2, we use 17 of these, along with the associated confidence scores that were extracted simultaneously using the OPENFACE toolkit. We investigate how the neural networks utilise FAUs so efficiently, what makes their learnings generalise across different cultures, and if

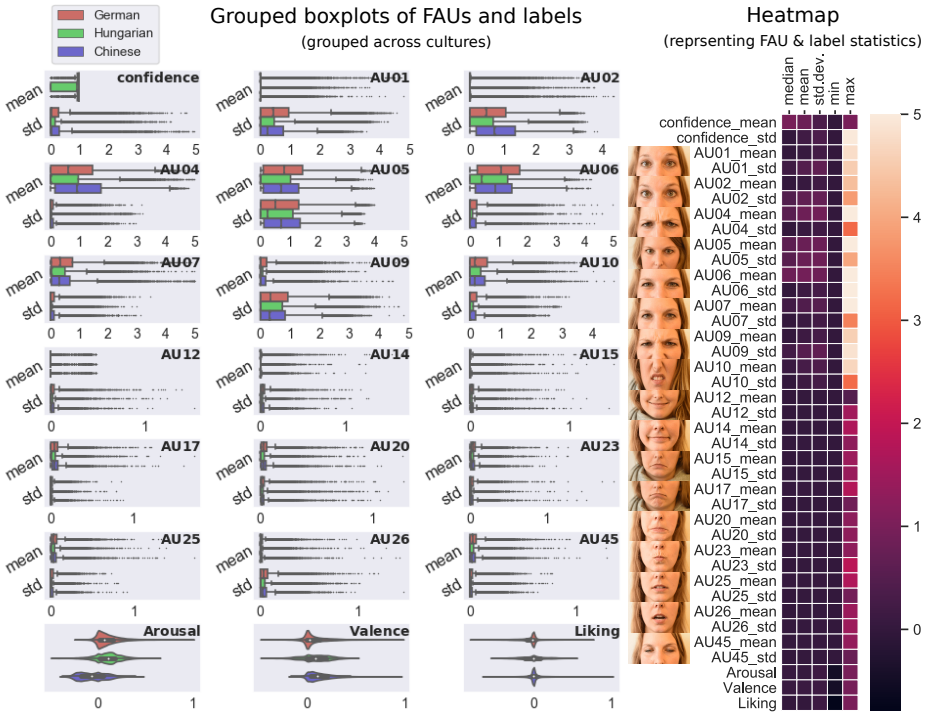

Figure 2: Boxplot of FAU feature activation levels and the arousal, valence, liking labels. As can be seen from the figure, the range (cf. x-axes) for different FAU features are drastically different. Consequently, a few FAU activations can dominate the rest of the features (cf. max values). On the other hand, a few other FAU features (e. g., AU12_mean) rarely ever manifest, even their maximum values are close to minimum of almost all other FAU feature activations.

these learnings are consistent to human perception of emotion expressions.

To explore the utility and challenges presented by the extracted FAU feature set, we first prepared statistical summary of the activation values of the individual features we would later train our models with (cf. Section 3.2). These input features are identical to the visual features provided to the participants of AVEC 2018 and AVEC 2019 challenges; i. e., the moving average and the moving standard deviations of 17 FAU activations computed for consecutive 50 frames (1 second) with a hop of 5 frames (100 ms). Section 3.2 also presents grouped box-plots of the corresponding moving mean and moving standard de-

viation values, for better comparison of these features across cultures. As can be seen from the plot, the ranges, quartiles, outlier distributions are similar across different cultures. Considering that the three cultures are vastly different from each another, the consistent statistics of FAU activations, across cultures – in some sense – again hint at the universality of the FAU features. One can expect, therefore, a superior pan-cultural predictive performance of the models when trained to predict emotions using the FAU features. In the discussions next, because the median utility function is less affected than the mean by the outliers in the data, we consider the median heatmap column to be better representatives of the features.

It can be inferred from the heatmap in Section 3.2 that not all FAU features convey equally useful information. For example, we notice that:

- The close to zero standard deviations ($\in [0.00, 0.07)$) of all of the AU12 to AU45 activation related features indicate that AU12 to AU45 related feature inputs to the model remain almost constant across the entire dataset. Medians of these features are close to zero likewise ($\in [0.00, 0.05]$), implying that the constant activation too is very close zero (< 0.05) likewise.

- The range for the moving-mean and moving-standard deviations are drastically different for different FAUs. For example, the maximum moving-mean activation for AU1 to AU10 FAUs is close to 5, while AU12 to AU45 moving-mean activations rarely go past 1.0 (cf. boxplots in Section 3.2).

- This dissimilarity between the different FAU features is so stark that even maximas (i. e., the outliers in one direction) of certain features (e. g., AU12_mean) are less than even medians of the most other features. Such extremely high or low values, in the absence of a batch-normalisation layer, can drive prediction of a neural network astray – unless the features with stronger activations are more informative already. It is only then that a model has a possibility to assign quasi-zero weights to the erratic features.

- Interestingly enough, in spite of the featured high maximas and close

to zero minimas (i. e., the high range) for AU4, AU6, AU7 and AU10-related features, the medians of the moving-standard deviations of activations are 0.000. Analysing further the time-series for moving-standard deviations of activations, we note that more than 50 % of the dataset is dominated by precisely zero-valued samples. It tells us that these four activations hardly ever fluctuate too much from their respective median values (.560, .835, .230, .100) in time.

- As for the remaining AU1, AU2, AU5 and AU9 activations, the medians of the moving-mean and the moving-standard deviations are (.000, .175), (.000, .440), (.603, .835), (.000, .235) respectively. As implied earlier, the first quantity represents roughly the degree of the FAU activation across the dataset in time. The second quantity, i. e., the median of standard deviation is a representative of the perturbation of that FAU activation level. The latter is affected by both the magnitude and the frequency of perturbation.

4. Experiment Design

We built and trained various minimalist neural network architectures using the Keras (v2.2.2) library with the Tensorflow (v1.11.0) backend. Training and evaluations are run on a regular notebook with an Nvidia GeForce GTX 1050 Ti GPU card. We perform the training on the full batch of FAU features obtained from 34 video sequences of German participants. We use RMSprop as the optimiser, with learning rate set to 0.001. We run the training for 2000 epochs, and choose the model weights based on its performance on the development set. Because the efforts to minimise the mean square error (MSE) do not necessarily translate into maximisation of concordance correlation coefficient (CCC) [92], we use deviation from the maximally achievable CCC i. e., '1-CCC' as the loss function of choice. This strategy, of employing CCC as part of the loss function, has proven to be successful in practice as well [79, 80, 93]. We choose the optimal delay compensation of 4.0 seconds and 2.8 seconds for arousal and valence

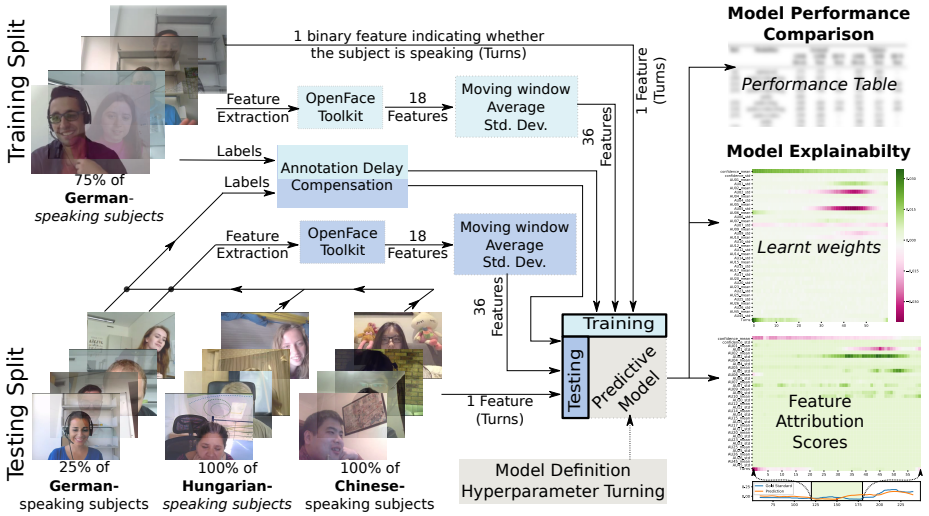

Figure 3: The entire experimental design pipeline

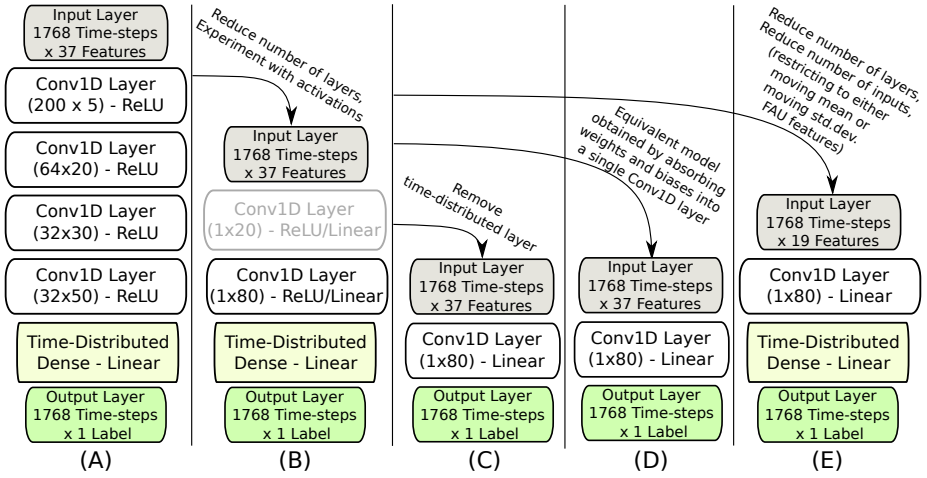

Figure 4: Different models we trained in our experiments, how we came up with a final shallow model.

respectively, based on our previous study [80].

4.1. Minimalist Model A

In our previous work, we demonstrated the suitability of CNNs over RNNs for a high noise, high resolution time-series prediction problems. We begin our

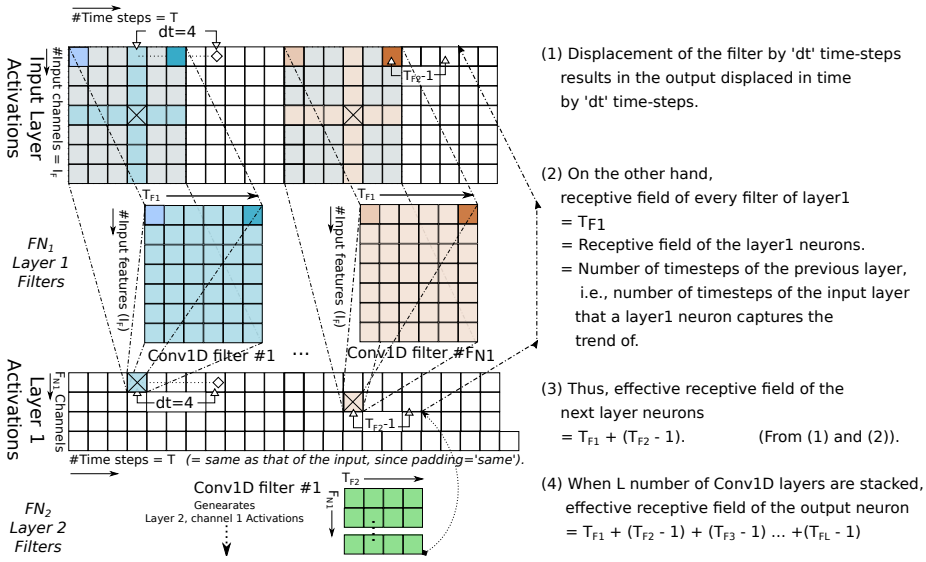

Figure 5: Receptive field of the output neuron when L number of Conv1D layers are stacked

$$= \left(\sum_{i=1}^{L} T_{Fi}\right) + 1 - L.$$

experiments with an identical CNN architecture, albeit we use visual features in place of the acoustic ones. Consequently, dimensionality of the input data reduces down to $N \times 1768 \times 37$ from $N \times 1768 \times 47$ (i. e., Num. Instances $\times$ Num. Time-steps $\times$ Num. Features). As a result, there is also reduction in the number of parameters the model needs to learn. A single output neuron in this architecture has a receptive field of about 10 seconds ($\because$ 50+30+20+5-3=102 time steps=10.2 seconds), as illustrated in Figure 5.

4.2. Minimalist Model B

Next, we simply removed the intermediate layers from the model A one by one, while keeping about the same receptive field for the output neurons. While we experimented with number of different combinations of number of intermediate layers (L_N), receptive field of the intermediate layers (T_F) and number of filers for each of the intermediate layer (N_F), we present in this study the results we obtained for the extreme case, where $N_F = 1$, $T_F =$

100. We experiment with different flavours of this model likewise; by changing the activation function of the only convolutional layer present in the model to 'linear'. Interestingly, the performance more or less, turned out to be similar for both these activations, as will be shown later in the section summarising all of the obtained results.

4.3. Minimalist Model C

The last layer of the Model B (cf. Figure 4) is a TimeDistributed Dense layer[5]. Because the output of the previous layer is $N \times 1768 \times 1$ matrix (as we use only 1 filter), the time-distributed layer only ends up scaling the outputs by a constant factor and adds a constant offset. We designed and trained a more reduced model architecture that is obtained by removing the TimeDistributed layer. However, this resulted in a deterioration of the model performance. This does not come as a surprise, as it is very common to witness the machine learning literature the post-processing of the outputs with constant offsets and constant scalings – as learnt from the development set. While a neural network is aimed at learning these very weights (scalings) and biases (offsets) directly, this ironic behaviour of neural networks – failing to learn and account for the constant scalings and offsets that otherwise would effectively reduce the overall loss – is witnessed yet again [93].

4.4. Minimalist Model D

We build a model with an input to output mapping identical to the trained model B, but of model C topology. If W_L^M and B_L^M are the weights and biases associated with the layer L of model M, and I and O denote inputs and outputs of the models, we have following input to output relationships.

[5] `https://keras.io/layers/wrappers/`

$$O = (W_{Conv1D}^{Model_B} * I + B_{Conv1D}^{Model_B}) * W_{Dense}^{Model_B} + B_{Dense}^{Model_B} \quad (1)$$

$$O = (W_{Conv1D}^{Model_D} * I + B_{Conv1D}^{Model_B}) \quad (2)$$

Comparing Equation (1) and Equation (2), we have:

$$W_{Conv1D}^{Model_D} = W_{Conv1D}^{Model_B} * W_{Dense}^{Model_B}, \quad (3)$$

and $$B_{Conv1D}^{Model_B} = B_{Conv1D}^{Model_B} * W_{Dense}^{Model_B} + B_{Dense}^{Model_B} \quad (4)$$

4.5. Minimalist Model E

In an attempt to make the model computationally even more inexpensive, we experimented with a few input feature combinations. These experiments too were data-driven, gaining insights from the visualisations of the filter weights and the feature attribution heatmaps, as we will discuss next.

5. Results and Insights

5.1. Feature Attribution Calculation

Consider model D consisting of only one intermediate Conv1D layer between the input and the output layer. Because the output is just a single channel (we predict only one emotion dimension at a time), number of filters associated with the intermediate Conv1D layer = 1. The output is computed by first multiplying the trained filter weights with a section of the input element-wise, summing the element-wise products with the trained bias value. Thus, the Conv1D layer essentially computes the degree of similarity of the input features (in an interval equal to its receptive field) against the pattern defined by the

weights of the trained filter.

Let the input to the model be defined by

$$X^{F\times T} = \{x_{f,t}\}, \qquad \text{where } f \in [0, F-1], \quad t \in [0, T-1],$$

and the corresponding output be defined by

$$Y^{1\times T} = \{y_t\}, \qquad \text{where } t \in [0, T-1].$$

Let the filter weights and bias be defined by the matrix

$$W^{F\times T_F} = \{w_{f,t}\}, \qquad \text{where } f \in [0, F-1], \quad t \in [0, T_F-1].$$

and scalar b respectively.

Therefore,

$$\begin{aligned} y_{t_0} &= g\left(\left[\sum_{f=0}^{F-1}\sum_{t=0}^{T_F-1} x_{f,(t_0-\frac{T_F}{2}+t)} \times w_{f,t}\right] + b\right) \text{where } g(\cdot) \text{ is an activation function.} \\ \therefore y_{t_0} &= \left[\sum_{f=0}^{F-1}\sum_{t=0}^{T_F-1} x_{f,(t_0-\frac{T_F}{2}+t)} \times w_{f,t}\right] + b \qquad \text{for the linear activation.} \end{aligned} \tag{5}$$

Dividing the constant bias b across all of the associated input features, feature attribution of the feature $x_{f,(t_0-\frac{T_F}{2}+t)}$ to the output y_{t_0} is $\left[x_{f,(t_0-\frac{T_F}{2}+t)} \times w_{f,t}\right] + \frac{b}{F\times T_F}$. We present insights we gained by inspecting these the heatmaps for these feature attribution scores for various time-steps (t_0 for arousal and valence labels, across several subjects. We present some of the plots in **??**.

5.2. Interpreting the Feature Attribution Matrix and the Filter Weights

There are several ways of interpreting the filter weights and the feature attribution plots. As discussed earlier, one interpretation of the filter weights is that the model looks for the pattern that is most similar to the one defined by the filter weight matrix overall. More similar the input pattern, higher is the output activation. Because output is ultimately the weighted linear combination of the filter coefficients (weighted by the input activations), the output can be grouped into several column-wise summations or several row-wise summations. We make following observations.

- For the strictly non-negative feature activations (as in the case of inputs in this study, the FAU activations), sign of a weight indicates whether a certain feature activation would end up contributing positively or negatively to the output.

- A dominant row – featuring high magnitude values, irrespective of the corresponding signs – indicates that the model amplifies the corresponding input feature heavily. It implies, thus, it is likely an important feature, and the model relies on this feature heavily when making the prediction. This interpretation, however, has to be taken with a pinch of salt in the case of non-normalised inputs, because the sheer magnitude of the weight does not directly translate to its importance. By assigning high magnitude weights, the model insures that the low input values are propagated through the model to drive the model output. A row of a filter weight matrix represents a temporal trend of the specific feature activation in the input matrix that the filter expects to see **??**(b). Effectively, the output is extent to which the input encapsulates the various expected temporal trends of the individual feature activations.

- Likewise, a column of a filter weight matrix represents a specific combination of all the feature values. When the provided inputs are FAU activations, an example combination could be 0.71 activation of lip corner puller (a smile), 0.51 cheek raiser FAU, and so on, mapping to arousal label of 0.83, through weighted summations **??**(c). Effectively, the output is extent to which the input encapsulates the various expected feature combinations at specific time-steps, relative to the centre of the filter.

- We note that the filter-weight alone can not quantify the relevance of individual features, since the scale of the input features is a crucial missing context. The feature attribution score tells us exactly to what extent a specific feature in time contributed to a specific output.

The very first feature relevance score matrices were very hard to interpret,

as these consisted of too many positive and negative scores. We added L_1 regularisation loss, effectively compelling the model to use a more sparse filter weight matrix. This results in a more sparse feature attribution matrix, which is lot easier to analyse for a human and to gain insights into. We visualise the matrices using heatmaps from the seaborn library [94], with a diverging PiYG colormap centred at 0.0, irrespective of the range of values it represents. While we experimented with the receptive field of the model D, the model was found to consistently rely on the following features: 1. Confidence_mean, 2. AU5_std 3. Turns.

6. Conclusion

References

[1] K. Courtney, et al., The use of social media in healthcare: organizational, clinical, and patient perspectives, Enabling health and healthcare through ICT: available, tailored and closer 183 (2013) 244.

[2] M. L. Antheunis, K. Tates, T. E. Nieboer, PatientsâĂŹ and health professionalsâĂŹ use of social media in health care: motives, barriers and expectations, Patient education and counseling 92 (3) (2013) 426–431.

[3] W. D. Evans, How social marketing works in health care, Bmj 332 (7551) (2006) 1207–1210.

[4] H. Cheng, P. Kotler, N. Lee, Social marketing for public health: global trends and success stories, Jones & Bartlett Learning, 2011.

[5] H. Korda, Z. Itani, Harnessing social media for health promotion and behavior change, Health promotion practice 14 (1) (2013) 15–23.

[6] R. Rozenblum, D. W. Bates, Patient-centred healthcare, social media and the internet: the perfect storm? (2013).

[7] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, L. Donaldson, Harnessing the cloud of patient experience: using social media to detect poor quality healthcare, BMJ Qual Saf 22 (3) (2013) 251–255.

[8] E. Smailhodzic, W. Hooijsma, A. Boonstra, D. J. Langley, Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals, BMC health services research 16 (1) (2016) 442.

[9] N. R. Armfield, L. C. Gray, A. C. Smith, Clinical use of skype: a review of the evidence base, Journal of Telemedicine and Telecare 18 (3) (2012) 125–127.

[10] D. B. Brecher, The use of skype in a community hospital inpatient palliative medicine consultation service, Journal of palliative medicine 16 (1) (2013) 110–112.

[11] S. Arik, J. Chen, K. Peng, W. Ping, Y. Zhou, Neural voice cloning with a few samples, in: Advances in Neural Information Processing Systems, 2018, pp. 10019–10029.

[12] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, in: Advances in neural information processing systems, 2018, pp. 4480–4490.

[13] L. Wan, Q. Wang, A. Papir, I. L. Moreno, Generalized end-to-end loss for speaker verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4879–4883.

[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499.

[15] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8).

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

[19] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.

[20] S. Amiriparian, M. Schmitt, S. Hantke, V. Pandit, B. Schuller, Humans Inside: Cooperative Big Multimedia Data Mining, in: A. Esposito, A. M. Esposito, L. C. Jain (Eds.), Innovations in Big Data Mining and Embedded Knowledge: Domestic and Social Context Challenges, Vol. 159 of Intelligent Systems Reference Library (ISRL), Springer, 2019, pp. 235–257, invited contribution.

[21] V. Pandit, S. Amiriparian, M. Schmitt, B. Schuller, Big Data Multimedia Mining: Feature Extraction facing Volume, Velocity, and Variety, in: S. Vrochidis, B. Huet, E. Y. Chang, I. Kompatsiaris (Eds.), Big Data Analytics for Large-Scale Multimedia Search, Wiley, 2019, Ch. 3, pp. 61–83.

[22] B. Borel, Clicks, lies and videotape, Scientific American 319 (4) (2018) 38–43.

[23] C. Esposito, A. De Santis, G. Tortora, H. Chang, K.-K. R. Choo, Blockchain: A panacea for healthcare cloud-based data security and privacy?, IEEE Cloud Computing 5 (1) (2018) 31–37.

[24] X. Yue, H. Wang, D. Jin, M. Li, W. Jiang, Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control, Journal of medical systems 40 (10) (2016) 218.

[25] K. Peterson, R. Deeduvanu, P. Kanjamala, K. Boles, A blockchain-based approach to health information exchange networks, in: Proc. NIST Workshop Blockchain Healthcare, Vol. 1, 2016, pp. 1–10.

[26] B. Malin, Betrayed by my shadow: learning data identity via trail matching, Journal of Privacy Technology 2147483647.

[27] B. Malin, L. Sweeney, How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems, Journal of biomedical informatics 37 (3) (2004) 179–192.

[28] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, H. Bischof, Mahalanobis distance learning for person re-identification, in: Person re-identification, Springer, 2014, pp. 247–267.

[29] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034.

[30] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.

[31] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806.

[32] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140.

[33] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences, arXiv preprint arXiv:1605.01713.

[34] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, in: International Conference on Artificial Neural Networks, Springer, 2016, pp. 63–71.

[35] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 3145–3153.

[36] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 3319–3328.

[37] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognition 65 (2017) 211–222.

[38] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint arXiv:1706.03825.

[39] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, S. Dähne, Learning how to explain neural networks: Patternnet and patternattribution, arXiv preprint arXiv:1705.05598.

[40] N. Liu, E.-S. Qi, M. Xu, B. Gao, G.-Q. Liu, A novel intelligent classification model for breast cancer diagnosis, Information Processing & Management 56 (3) (2019) 609–623.

[41] C. Janott, M. Schmitt, Y. Zhang, K. Qian, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, B. Schuller, Snoring Classified: The Munich Passau Snore Sound Corpus, Computers in Biology and Medicine 94 (1) (2018) 106–118, (IF: 2.115 (2017)).

[42] K. Qian, C. Janott, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, B. Schuller, Classification of the Excitation Location of Snore Sounds in the Upper Airway by Acoustic Multi-Feature Analysis, IEEE Transactions on Biomedical Engineering 64 (8) (2017) 1731–1741, (IF: 4.288 (2017)).

[43] T. Curtis, T. Gardiner, A. Stitt, Microvascular lesions of diabetic retinopathy: clues towards understanding pathogenesis?, Eye 23 (7) (2009) 1496.

[44] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy, Procedia Computer Science 90 (2016) 200–205.

[45] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[46] M. Thelwall, Tensistrength: Stress and relaxation magnitude detection for social media texts, Information Processing & Management 53 (1) (2017) 106–121.

[47] M. Yoo, S. Lee, T. Ha, Semantic network analysis for understanding user experiences of bipolar and depressive disorders on reddit, Information Processing & Management 56 (4) (2019) 1565–1575.

[48] S. Amiriparian, A. Awad, M. Gerczuk, L. Stappen, A. Baird, S. Ottl, B. Schuller, Audio-based Recognition of Bipolar Disorder Utilising Capsule

Networks, in: Proceedings 32nd International Joint Conference on Neural Networks (IJCNN), INNS/IEEE, IEEE, Budapest, Hungary, 2019, pp. 1–7.

[49] L. Roche, D. Zhang, F. B. Pokorny, B. W. Schuller, G. Esposito, S. Bölte, H. Roeyers, L. Poustka, K. D. Bartl-Pokorny, M. Gugatschka, H. Waddington, R. Vollmann, C. Einspieler, P. B. Marschik, Early Vocal Development in Autism Spectrum Disorders, Rett Syndrome, and Fragile X Syndrome: Insights from Studies using Retrospective Video Analysis, Advances in Neurodevelopmental Disorders 2 (1) (2018) 49–61.

[50] B. Schuller, Can Virtual Human Interviewers "Hear" Real Humans' Depression?, IEEE Computer Magazine 49 (7) (2016) 8, (IF: 1.940 (2017)).

[51] K. R. Scherer, G. Ceschi, Lost luggage: a field study of emotion–antecedent appraisal, Motivation and emotion 21 (3) (1997) 211–235.

[52] F. Schiel, S. Steininger, U. Türk, The SmartKom Multimodal Corpus at BAS., in: LREC, 2002.

[53] M. Grimm, K. Kroschel, S. Narayanan, The vera am mittag german audio-visual emotional speech database, in: ICME, IEEE, 2008, pp. 865–868.

[54] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schroder, The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent, IEEE Transactions on Affective Computing 3 (1) (2012) 5–17.

[55] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, AVEC 2013 âĂŞ- The Continuous Audio/Visual Emotion and Depression Recognition Challenge, in: Proc. 3rd ACM international workshop on Audio/Visual Emotion Challenge, ACM, ACM, Barcelona, Spain, 2013, pp. 3–10.

[56] I. Sneddon, M. McRorie, G. McKeown, J. Hanratty, The belfast induced natural emotion database, IEEE Transactions on Affective Computing 3 (1) (2012) 32–41.

[57] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendeventer, D. W. Cunningham, C. Wallraven, Cardiff conversation database (ccdb): A database of natural dyadic conversations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 277–282.

[58] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–8.

[59] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, AVEC 2014 – The Three Dimensional Affect and Depression Challenge, in: Proc. 4th ACM international workshop on Audio/Visual Emotion Challenge, ACM, ACM, Orlando, FL, 2014.

[60] S. Bilakhia, S. Petridis, A. Nijholt, M. Pantic, The mahnob mimicry database: A database of naturalistic human interactions, Pattern recognition letters 66 (2015) 52–61.

[61] J. Vandeventer, A. J. Aubrey, P. L. Rosin, D. Marshall, 4D Cardiff Conversation Database (4D CCDb): A 4D database of natural, dyadic conversations, in: Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015), 2015.

[62] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, E. Hajiyev, M. Pantic, SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild, IEEE Transactions on Pattern Analysis and Machine Intelligence 41.

[63] V. Pandit, N. Cummins, M. Schmitt, S. Hantke, F. Graf, L. Paletta, B. Schuller, Tracking Authentic and In-the-wild Emotions using Speech, in: Proc. 1st ACII Asia 2018, AAAC, IEEE, Beijing, P. R. China, 2018.

[64] F. Eyben, F. Weninger, L. Paletta, B. Schuller, The acoustics of eye contact – Detecting visual attention from conversational audio cues, in: Proc. 6th

Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction (GAZEIN) at 15th ICMI, ACM, Sydney, Australia, 2013, pp. 7–12.

[65] V. Pandit, M. Schmitt, N. Cummins, F. Graf, L. Paletta, B. Schuller, How Good Is Your Model 'Really'? On 'Wildness' of the In-the-wild Speech-based Affect Recognisers, in: Proc. 20th Intl. Conf. Speech and Computer, SPECOM 2018, ISCA, Springer, Leipzig, Germany, 2018.

[66] F. Ringeval, B. Schuller, M. Valstar, S. Mozgai, N. Cummins, M. Schmitt, M. Pantic, AVEC 2017 – Real-life Depression, and Affect Recognition Workshop and Challenge, in: Proc. 7th Int. Workshop on Audio/Visual Emotion Challenge (AVEC'17) at 25th ACM MM, ACM, Mountain View, CA, 2017, pp. 3–9.

[67] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftci, H. Gülec, A. A. Salah, M. Pantic, AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition, in: Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018, ACM, ACM, Seoul, South Korea, 2018.

[68] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, M. Soleymani, M. Schmitt, S. Amiriparian, E.-M. Messner, L. Tavabi, S. Song, S. Alisamir, S. Lui, Z. Zhao, M. Pantic, AVEC 2019 Workshop and Challenge: State-of-Mind, Depression with AI, and Cross-Cultural Affect Recognition, in: F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, M. Pantic (Eds.), Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge, AVEC'19, co-located with the 27th ACM International Conference on Multimedia, MM 2019, ACM, ACM, Niece, France, 2018.

[69] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, J. Yi, Continuous multimodal emotion prediction based on long short term memory recurrent neu-

ral network, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, 2017, pp. 11–18.

[70] S. Chen, Q. Jin, J. Zhao, S. Wang, Multimodal multi-task learning for dimensional and continuous emotion recognition, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, 2017, pp. 19–26.

[71] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, J. Epps, Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, 2017, pp. 27–35.

[72] K. Wataraka Gamage, T. Dang, V. Sethu, J. Epps, E. Ambikairajah, Speech-based continuous emotion prediction by learning perception responses related to salient events: A study based on vocal affect bursts and cross-cultural affect in avec 2018, in: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, ACM, 2018, pp. 47–55.

[73] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, M. Yang, Multimodal continuous emotion recognition with data augmentation using recurrent neural networks, in: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, ACM, 2018, pp. 57–64.

[74] J. Zhao, R. Li, S. Chen, Q. Jin, Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions, in: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, ACM, 2018, pp. 65–72.

[75] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, H. Sahli, Efficient spatial temporal convolutional features for audiovisual continuous affect recognition, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, ACM, 2019, pp. 19–26.

[76] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov, A. A. Salah, Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, ACM, 2019, pp. 27–35.

[77] J. Zhao, R. Li, J. Liang, S. Chen, Q. Jin, Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, ACM, 2019, pp. 37–45.

[78] M. Schmitt, B. Schuller, openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit, J. Mach. Learn. Res. 18.

[79] V. Pandit, M. Schmitt, N. Cummins, B. Schuller, I know how you feel now, and here's why!: Demystifying Time-continuous High Resolution Text-based Affect Predictions In the Wild, in: Proceedings of the 32nd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2019, IEEE, IEEE, Córdoba, Spain, 2019, pp. 465–470.

[80] M. Schmitt, N. Cummins, B. W. Schuller, Continuous Emotion Recognition in Speech – Do We Need Recurrence?, in: Proceedings INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association, ISCA, ISCA, Graz, Austria, 2019, 5 pages, to appear (acceptance rate: 49.3 %).

[81] Y. Bengio, P. Simard, P. Frasconi, et al., Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks 5 (2) (1994) 157–166.

[82] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[83] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares,

H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078.

[84] R. Jozefowicz, W. Zaremba, I. Sutskever, An empirical exploration of recurrent network architectures, in: International Conference on Machine Learning, 2015, pp. 2342–2350.

[85] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555.

[86] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: International conference on machine learning, 2013, pp. 1310–1318.

[87] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[88] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, K. Truong, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202.

[89] F. Eyben, M. Wöllmer, B. Schuller, openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor, in: Proc. 18th ACM MM, ACM, Florence, Italy, 2010, pp. 1459–1462.

[90] F. Eyben, F. Weninger, F. Groß, B. Schuller, Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor, in: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, ACM, ACM, Barcelona, Spain, 2013, pp. 835–838, (Honorable Mention (2nd place) in the ACM MM 2013 Open-source Software Competition, acceptance rate: 28 %).

[91] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L.-P. Morency, Openface 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 59–66.

[92] V. Pandit, B. Schuller, On Many-To-Many Mappings Between Concordance Correlation Coefficient and Mean Square Error, arXiv preprint:1902.05180.

[93] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, S. Zafeiriou, Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network, in: Proc. 41st ICASSP, IEEE, Shanghai, P. R. China, 2016, pp. 5200–5204.

[94] M. Waskom, O. Botvinnik, P. Hobson, J. B. Cole, Y. Halchenko, S. Hoyer, A. Miles, T. Augspurger, T. Yarkoni, T. Megies, et al., seaborn: v0. 5.0 (november 2014), Zenodo, doi 10.